Group 3: Jeannie He, Li Cheng, Yifei Dong, Yusen Wang

H.W.2.1.a)

f is Lipschitz continuous iff $\|w\|_2 \leq D \Rightarrow \|\nabla f(w)\|_2 \leq B$

We have $f(w) = \frac{1}{N}\sum_{i\in[N]} f_i(w) + \lambda \|w\|_2^2$ where $f_i(w) = \log(1+\exp\{-y_i w^T x_i\})$.

Since $\nabla f_i(w) = \frac{\exp\{-y_i w^T x_i\}}{1+\exp\{-y_i w^T x_i\}}(-y_i x_i) = \frac{-y_i x_i}{1+\exp\{y_i w^T x_i\}}$, we have:

$$\|\nabla f(w)\|_2 = \left\|\frac{1}{N}\sum_{i\in[N]} \nabla f_i(w) + 2\lambda w\right\|_2 = \left\|\frac{1}{N}\sum_{i\in[N]} \frac{-y_i x_i}{\underbrace{1+\exp\{y_i w^T x_i\}}_{\geq 0}} + 2\lambda w\right\|_2$$

$$\leq \frac{1}{N}\left\|\sum_{i\in[N]} y_i x_i\right\|_2 + 2\lambda \|w\|_2$$

Hence, if $\|w\|_2 \leq D$, then $\|\nabla f(w)\|_2 \leq \frac{1}{N}\left\|\sum_{i\in[N]} y_i x_i\right\|_2 + 2\lambda D$

This shows that f is Lipschitz continuous with constant

$$B = \frac{1}{N}\left\|\sum_{i\in[N]} y_i x_i\right\|_2 + 2\lambda D .$$

H.W. 2.1 b)

We have from Lecture 3 that:

$f_i(w) = \log(1+\exp(-y_i w^T x_i))$ is L-smooth iff $f_i(w_2) \leq f_i(w_1) + \nabla f_i(w_1)^T (w_2 - w_1) + \frac{L}{2}\|w_2 - w_1\|_2^2$

From lecture 2 & assignment 1.2a), we have that this holds if: $\|\nabla f(w_2) - \nabla f(w_1)\|_2 \leq L\|w_2 - w_1\|_2$

We have: $\nabla f_i(w) = \frac{\exp\{-y_i w^T x_i\}}{1+\exp\{-y_i w^T x_i\}}(-y_i x_i) = \frac{(-y_i x_i)}{1+\exp\{y_i w^T x_i\}}$

$\Rightarrow \|\nabla f_i(w_2) - \nabla f_i(w_1)\|_2 = \left\| \frac{(-y_i x_i)}{1+\exp\{y_i w_2^T x_i\}} - \frac{(-y_i x_i)}{1+\exp\{y_i w_1^T x_i\}} \right\|_2$ (2.1.b.i)

$= \|y_i x_i\|_2 \left\| \frac{1}{1+\exp\{y_i w_2^T x\}} - \frac{1}{1+\exp\{y_i w_1^T x_i\}} \right\|_2$

Since $0 \leq \frac{1}{1+\exp\{a\}} \leq 1$ for $a \in \mathbb{R}$, we have $\left\| \frac{1}{1+\exp\{b\}} - \frac{1}{1+\exp\{a\}} \right\|_2 \leq \|1-0\|_2 \leq 1$ for $a, b \in \mathbb{R}$

Hence, (2.1.b.i) gives: $\|\nabla f_i(w_2) - \nabla f_i(w_1)\|_2 \leq \|y_i x_i\|_2$ (2.1.b.ii)

Hence, $f_i(w)$ is L-smooth with $L = \frac{\|y_i x_i\|_2}{\|w_2 - w_1\|_2}$

Since $f(w) = \frac{1}{N}\sum_{i\in[N]} f_i(w) + \lambda\|w\|_2^2 \Rightarrow \nabla f(w) = \frac{1}{N}\sum_{i\in[N]} \nabla f_i(w) + 2\lambda w$

We have: $\|\nabla f(w_2) - \nabla f(w_1)\|_2 = \left\| \left(\frac{1}{N}\sum_{i\in[N]} \nabla f_i(w_2) + \lambda w_2\right) - \left(\frac{1}{N}\sum_{i\in[N]} \nabla f_i(w_1) + \lambda w_1\right) \right\|_2$

$\leq \frac{1}{N}\sum_{i\in[N]} \|\nabla f_i(w_2) - \nabla f_i(w_1)\|_2 + |\lambda|\,\|w_2 - w_1\|_2 \leq \{2.1.b.ii\} \leq \frac{1}{N}\sum_{i\in[N]} \|y_i x_i\|_2 + |\lambda|\,\|w_2 - w_1\|_2$

Hence, $f(w)$ is L-smooth with $L = \frac{\frac{1}{N}\sum_{i\in[N]} \|y_i x_i\|_2}{\|w_2 - w_1\|_2} + |\lambda|$

should be positive

H.W. 2.1 c)

f is strongly convex with constant $\mu > 0$ iff

$f(w_2) \geq f(w_1) + \nabla f(w_1)^T (w_2 - w_1) + \frac{\mu}{2}\|w_2 - w_1\|_2^2$, which is equivalent to:

$g(w) = f(w) - \frac{\mu}{2}\|w\|^2$ is convex because of the 1st order condition for convexity

Following the monotone gradient condition for convexity, we have therefore that $f(w)$ is strongly convex iff $(\nabla f(w_2) - \mu w_1 - \nabla f(w_1) - \mu w_2)^T (w_2 - w_1) \geq 0$,

Which is equivalent to $(\nabla f(w_2) - \nabla f(w_1))^T (w_2 - w_1) \geq \mu \|w_2 - w_1\|_2^2$

We have: $\nabla f_i(w) = \frac{\exp\{-y_i w^T x_i\}}{1+\exp\{-y_i w^T x_i\}}(-y_i x_i) = \frac{-y_i x_i}{1+\exp\{y_i w^T x_i\}}$

and $\nabla f(w) = \frac{1}{N}\sum_{i\in[N]} \nabla f_i(w) + 2\lambda w$

$$(\nabla f(w_2) - \nabla f(w_1))^T (w_2 - w_1) = \left[\left(\frac{1}{N}\sum_{i\in[N]} \nabla f_i(w_2) + 2\lambda w_2\right) - \left(\frac{1}{N}\sum_{i\in[N]} \nabla f_i(w_1) + 2\lambda w_1\right)\right]^T (w_2 - w_1) =$$

$$\cancel{= \left(\frac{1}{N}\sum_{i\in[N]} \nabla f_i(w_2) - \nabla f_i(w_1)\right)^T (w_2 - w_1) + 2\lambda (w_2 - w_1)^T (w_2 - w_1) =}$$

$$= \left(\frac{1}{N}\sum_{i\in[N]} \frac{-y_i x_i}{1+\exp\{y_i w_2^T x_i\}} - \frac{-y_i x_i}{1+\exp\{y_i w_1^T x_i\}}\right)^T (w_2 - w_1) + 2\lambda \|w_2 - w_1\|_2^2 =$$

$$= \left(\frac{1}{N}\sum_{i\in[N]} \frac{y_i x_i}{1+\exp\{y_i w_1^T x_i\}} - \frac{y_i x_i}{1+\exp\{y_i w_2^T x_i\}}\right)^T (w_2 - w_1) + 2\lambda \|w_2 - w_1\|_2^2 =$$

$$= \left(\frac{1}{N}\sum_{i\in[N]} \frac{y_i x_i \left(\exp\{y_i w_2^T x_i\} - \exp\{y_i w_1^T x_i\}\right)}{(1+\exp\{y_i w_1^T x_i\})(1+\exp\{y_i w_2^T x_i\})}\right)^T (w_2 - w_1) + 2\lambda \|w_2 - w_1\|_2^2 \geq$$

$$\geq 0 + 2\lambda \|w_2 - w_1\|_2^2 = 2\lambda \|w_2 - w_1\|_2^2$$

Hence, $f(w)$ is strongly convex with $\mu = 2\lambda$.

H.W. 2.2.

Since $Var[X] = \mathbb{E}[X^2] - \mathbb{E}[X]^2 \Leftrightarrow \mathbb{E}[X^2] = Var[X] + \mathbb{E}[X]^2$,

We have:

$$\mathbb{E}_{\xi_k}[\|g(w_k;\xi_k)\|^2] = \|\mathbb{E}_{\xi_k}[g(w_k;\xi_k)]\|_2^2 + Var_{\xi_k}[g(w_k;\xi_k)] \qquad (2.2.i)$$

Since $0 \leq \|\mathbb{E}_{\xi_k}[g(w_k;\xi_k)]\|_2 \leq C_0\|\nabla f(w_k)\|_2$, we have:

$$\|\mathbb{E}_{\xi_k}[g(w_k;\xi_k)]\|_2^2 \leq C_0\|\nabla f(w_k)\|_2^2. \qquad (2.2.ii)$$

Substituting (2.2.ii) and $Var_{\xi_k}[g(w_k;\xi_k)] \leq M + M_V\|\nabla f(w_k)\|_2^2$ into (2.2.i) gives:

$$\mathbb{E}_{\xi_k}[\|g(w_k;\xi_k)\|^2] \leq C_0\|\nabla f(w_k)\|_2^2 + M + M_V\|\nabla f(w_k)\|_2^2 =$$

$$= M + (C_0 + M_V)\|\nabla f(w_k)\|_2^2$$

We have thus proved that $\mathbb{E}_{\xi_k}[\|g(w_k;\xi_k)\|^2] \leq \alpha + \beta\|\nabla f(w_k)\|_2^2$

and found that $\alpha = M$, $\beta = C_0 + M_V$. Q.E.D.

HW2.3 For SGD with non-convex objective functions. Prove that with square summable but not summable step-size, for any $k \in \mathbb{N}$,

$$E\left[\sum_{k \in [k]} \alpha_k \|\nabla f(w_k)\|_2^2\right] < \infty.$$

and therefore, $E\left[\frac{1}{\sum_{k \in [k]} \alpha_k} \sum_{k \in [k]} \alpha_k \|\nabla f(w_k)\|_2^2\right] \xrightarrow{k \to \infty} 0$

Proof:

Generic SG on L-smooth function satisfies:

$$E[f(w_{k+1})] - f(w_k) \leq -\left(c - \frac{1}{2}\alpha_k L M_G\right)\alpha_k \|\nabla f(w_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L M. \quad (1)$$

For $k \in [k]$, the total expectation of (1) is:

$$f_{inf} - f(w_1) \leq E[f(w_{k+1})] - f(w_1)$$
$$\leq -cE\left[\sum_{k \in [k]} \alpha_k \|\nabla f(w_k)\|^2\right] + \frac{1}{2} L M_G E\left[\sum_{k \in [k]} \alpha_k^2 \|\nabla f(w_k)\|^2\right] + \frac{LM}{2} \sum_{k \in [k]} \alpha_k^2$$

Then,

$$E\left[\sum_{k \in [k]} \alpha_k \|\nabla f(w_k)\|^2\right] \leq \frac{f(w_1) - f_{inf}}{c} + \frac{L M_G}{2c} E\left[\sum_{k \in [k]} \alpha_k^2 \|\nabla f(w_k)\|^2\right] + \frac{LM}{2c} \sum_k \alpha_k^2 \quad (2)$$

Since f is L-smooth, f must be Lipschitz continuous,

Then $\|\nabla f(w_k)\|$ is bounded on $\{w_k : \|w_k\| \leq D\}$. i.e., $\|\nabla f(w_k)\| \leq B$.

Hence,

$$E\left[\sum_{k \in [k]} \alpha_k^2 \|\nabla f(w_k)\|^2\right] \leq B^2 E\left[\sum_{k \in [k]} \alpha_k^2\right] < \infty.$$

(Since $\sum_{k \in [k]} \alpha_k^2 < \infty$)

Since in (2), $\frac{f(w_1) - f_{inf}}{c}$, $\frac{L M_G}{2c} E\left[\sum_{k \in [k]} \alpha_k^2 \|\nabla f(w_k)\|^2\right]$, $\frac{LM}{2c} \sum_k \alpha_k^2$ are all bounded.

Then we have $E\left[\sum_{k \in [k]} \alpha_k \|\nabla f(w_k)\|^2\right] < \infty$ ∎.

$$\mathbb{E}\left[\frac{1}{\sum \alpha_k} \sum \alpha_k \|\nabla f(w_k)\|^2\right] \le \frac{f(w_1) - f_{inf}}{c \cdot \sum \alpha_k} + \frac{2M_G B^2 + LM}{2c} \cdot \frac{\sum \alpha_k^2}{\sum \alpha_k}$$

sine $\lim_{k \to \infty} \sum_{k \in [k]} \alpha_k = \infty$ and $\lim_{k \to \infty} \sum_{k \in [k]} \alpha_k^2 < \infty$,

it holds that,

$$\frac{1}{\sum \alpha_k} \xrightarrow{k \to \infty} 0 \quad \text{and} \quad \frac{\sum \alpha_k^2}{\sum \alpha_k} \longrightarrow 0$$

Therefore,

$$\mathbb{E}\left[\frac{1}{\sum \alpha_k} \sum \alpha_k \|\nabla f(w_k)\|^2\right] \xrightarrow{k \to \infty} 0$$